Noname manuscript No.
(will be inserted by the editor)

Discussion of "Multivariate Functional Outlier Detection" by M. Hubert, P. Rousseeuw and P. Segaert

Alicia Nieto-Reyes · Juan A. Cuesta-Albertos

Received: date / Accepted: date

1 Introduction

Firstly we congratulate the authors on a wonderful paper full of new nice ideas, which can be considered as a major breakthrough in the functional outlier detection using visual procedures. These ideas include a taxonomy of outliers, the definition of bag distance and the centrality-stability plots. Indeed, each of the last two ideas is fundamental in the development of each of the two corresponding procedures to detect multivariate functional outliers introduced in the paper. The first procedure consists in constructing a heat map using the functional bag distance based on the Tukey depth while the second in a scatter-plot based on the skew-adjusted projection depth, SPD, named the centrality-stability plot, CSP.

Moreover, both procedures complement themselves because the heat maps are good in detecting all kind of outliers excepting the shape outliers, but those are clearly identified with the CSP's

Our discussion focus, firstly, on shedding light on the behaviour of the proposed procedures when applied to multivariate functional data whose dimension is entitled to be extremely high. Secondly, a simplification of the CSP is proposed. Furthermore, we encourage the authors to comment on the advantages/disadvantages of applying, what they call, the MFSPD versus, what they call, the $1/(1 + \text{FAO})$, as the difference between them just lies in the reverse order of the integral and inverse functionals.

Research partially supported by the Spanish Ministerio de Ciencia y Tecnología, grant MTM2011-28657-C02-02.

Departamento de Matemáticas, Estadística y Computación, Universidad de Cantabria.
Avda. los Castros s.n., 39005 Santander, Spain
E-mail: alicia.nieto@unican.es

Departamento de Matemáticas, Estadística y Computación, Universidad de Cantabria.
Avda. los Castros s.n., 39005 Santander, Spain

2 High dimensional functional data

The ideas used in the procedures proposed in [4] apply, in principle, to multivariate functional data of any dimension. However, those procedures rely in techniques that make them being only applicable to multivariate functional data of low dimension. The reason for this is that, as commented in the introduction, the first procedure is based on the multivariate Tukey depth and the second on the SPD; and exact computations of these depths are only possible in multivariate spaces of low dimension [6,5]. Note that these methodologies are applied in [4] to multivariate functional datasets of only dimension 1 to 3.

The random Tukey depth [1] is a statistical depth that approximates the Tukey depth. According to [1,3] it needs of a very low number of projections to obtain equivalent results to those of the Tukey depth. Due to this fact, the random Tukey depth is very fast to compute, making it the depth to go for, not only when the dimension of the space is moderate or high, but also when it is low due to its computationally effectiveness. Additionally, the random Tukey depth inherits from the Tukey depth the nice properties that made it well-known [1,2]. Consequently, we propose a modification of the first procedure in [4] consisting in basing it on the random Tukey depth.

Concerning the second procedure, the authors acknowledge the difficulty in the computation of the SPD and they propose to compute it in dimension p, by taking $250p$ sets of $p+1$ randomly chosen points and using the orthogonal directions to the affine hyperplanes those sets determine. We believe that the same idea we employed for the random Tukey depth should also work in this setting, but we do not develop it here as the theory underneath the random projection depth has to yet be studied, and this is well beyond the scope of this discussion.

Through the rest of the section we show how the first procedure based on the random Tukey depth works as the one proposed in [4], but with the advantage of being effective in any dimension. Indeed, we apply it here to data in dimension 200 in contrast to just in dimension 1, as it is done in [4]. This type of data appear, for instance, in the analysis of some brain disorders, where 200 (or even more) sensors measure simultaneously the activity of 200 points in the brain; and the resulting data in each of those points is a curve. Precisely, a goal of these studies is to discover sensors showing an anomalous behavior; i.e. to discover outliers.

A comparison with the results obtained in [4] is not possible because there the authors only use the Tukey depth with one-dimensional curves, in this case, the Tukey and the random Tukey depths, obviously, coincide. However, taking into account that the dataset used in in [4] consists of 39 curves measured at 226 time points, we have also simulated 39 curves measured at 226 time points but in dimension 200 instead of in dimension 1.

To simulate the data we use for each of the dimensions the methodology proposed for Gaussian processes in [7, Section 2.2]. The first 12 dimensions, out of 200, of the simulated data are plotted in Figure 1. Thus, we obtain independent dimensions; of course, it is of great interest to study this problem

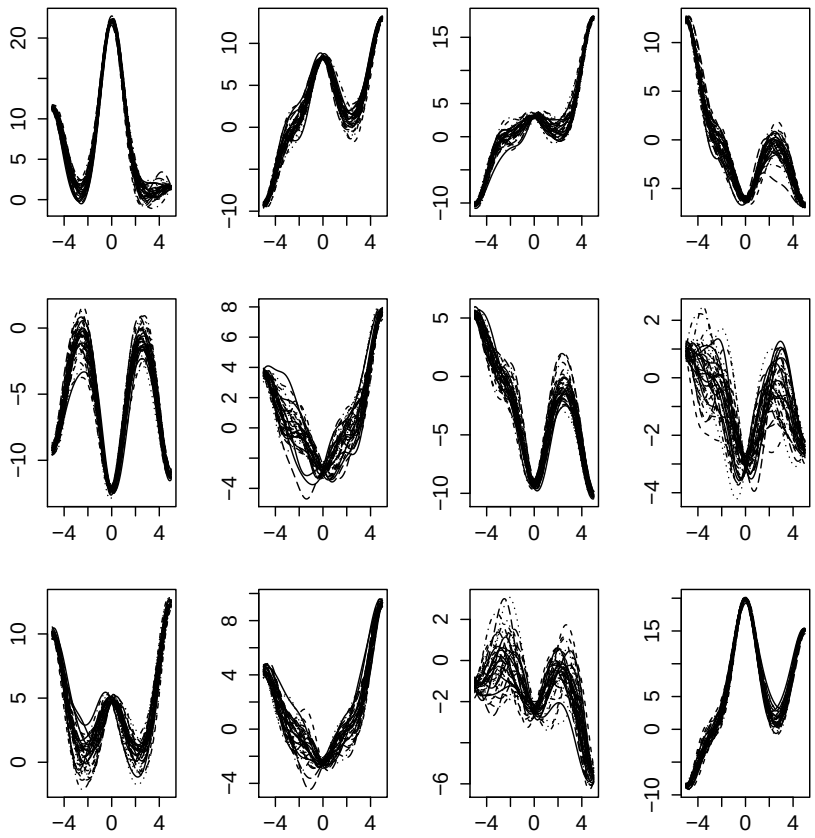

Fig. 1 First 12 dimensions, from left to right and top to bottom, of the simulated data before including outliers.

under simulated data with a more complex covariance structure across the dimensions; however, we do not consider it here to ease the explanation on the construction of the outliers and not extend the length of this discussion.

Seven outliers, containing the whole variety provided in the taxonomy of [4], are included in the simulated data. For this aim, the curves are labelled from 1 to 39; being the outliers the curves 1 to 7. Particularly, curves 1 to 4 are shift outliers, curve 5 an isolated outlier, curve 6 an amplitude outlier and curve 7 a shape outlier. In detail:

1. Curve 1 is a shift outlier only in dimension 1 over the whole time range.
2. Curve 2 is a shift outlier in every dimension and over the whole time range.
3. Curve 3 is a shift outlier in dimension 3 and over the first half of the time range.

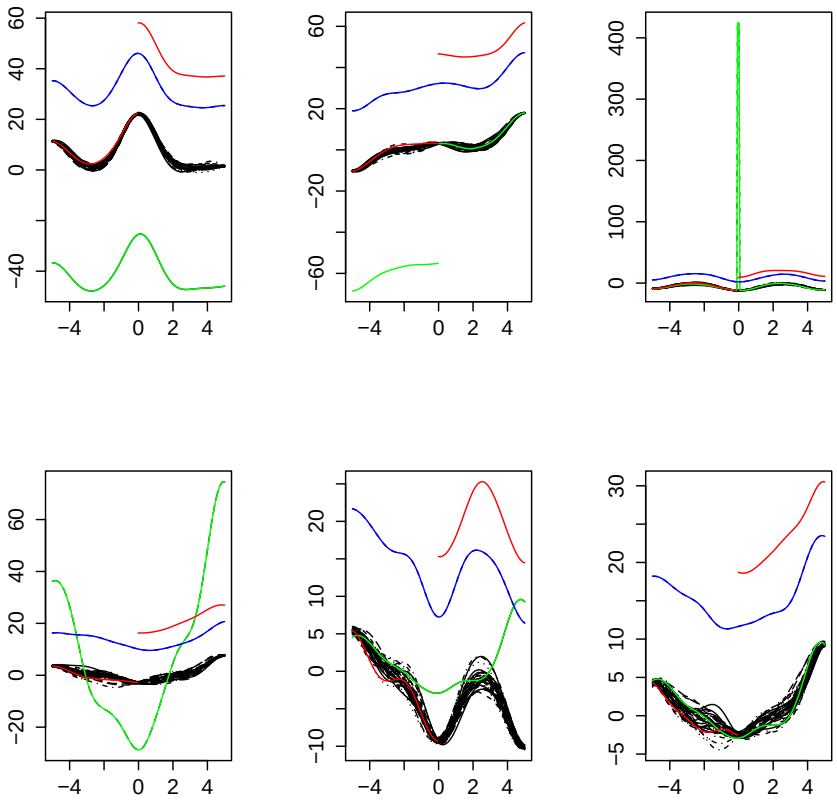

Fig. 2 Dimensions 1, 3, 5, 6, 7 and 10 (left to right and top to bottom) of the simulated data after including outliers. The colours are explained in the text.

4. Curve 4 is a shift outlier in every dimension and over the second half of the time range.
5. Curve 5 is an isolated outlier in dimension 5 at the three points in the middle of the time range.
6. Curve 6 is an amplitude outlier in dimension 6 over the whole time range.
7. In dimension 7, Curve 7 is a shape outlier over the whole time range and a shift outlier over part of the time range. However, in dimension 7, this curve is less distant to the curves cloud than the rest of, shift, outliers in this dimension: Curve 2 and Curve 4. It is constructed by substituting the original Curve 7 in dimension 7 by the Curve 7 in dimension 10.

To appreciate the outliers, we have plotted in Figure 2 the data in dimensions 1, 3, 5, 6, 7 and 10 from left to right and top to bottom. Curve 2 is in blue

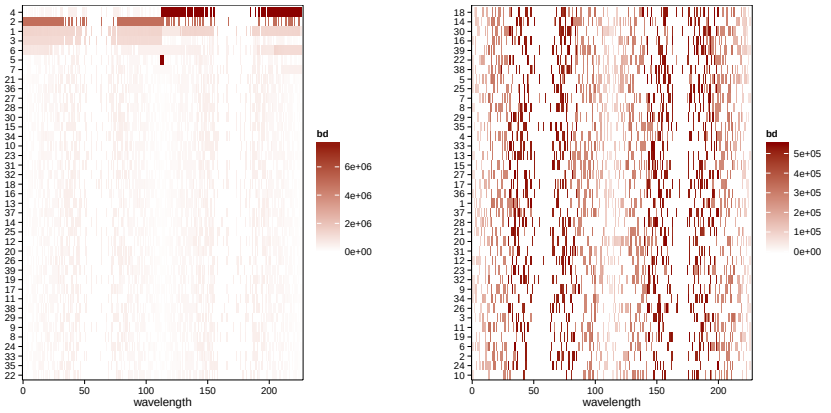

Fig. 3 Heat map of the simulated data after including the outliers (left) and before including the outliers (right).

color and Curve 4 in red in all plots. From left to right and top to bottom, the green curve represents the Curve 1 (respectively, 3, 5, 6, 7 and 7) in dimension 1 (resp. 3, 5, 6, 7, and 10). Notice that Curve 7 in dimension 10 is not an outlier because this curve is the same in dimensions 7 and 10, because Curve 7 in dimension 10 has been copied to dimension 7 to have a shape outlier in the former dimension.

The first procedure proposed in [4] is applied using the random Tukey depth and it is obtained the heat map shown in the left plot of Figure 3. The random Tukey depth is applied here using only 20 random projections. These random projections are drawn with the uniform distribution on the sphere, as suggested in [1]. This makes the procedure computationally effective while providing the desirable results, i.e. spotting the outliers. However, the procedure is still computationally effective if we take 10 times more projections, or even more. Nevertheless, it suffices with a low number of projections to get the appropriate results.

In the left plot of Figure 3 we can appreciate that the first six outliers are easily spotted. Particularly, it can be clearly appreciated that Curve 5 is an isolated outlier in the middle of the time range or that Curve 4 is an outlier only over the second half of the time range.

Concerning outlier number 7, as stated, heat maps are not intended to detect shape outliers and, then, this curves is unnoticed excepting for a very slightly dark color in the upper part of the wavelengths.

It is worth commenting that the white vertical areas in the heat map in the left plot of Figure 3 are a particularity of the data we have simulated. To make this clear we have plotted in the right plot of Figure 3 the heat map of the simulated data before including the outliers.

3 A simplification of the centrality-stability plots

Among the contents of the paper, we have specially enjoyed the part devoted to the centrality-stability plots. The basic idea behind the CSP's is that given a family of curves $Y_j, j = 1, \ldots, n$ which, in order to simplify the exposition, we assume measured at the same time points $t_j \in [a, b], j = 1, \ldots, T$, the CSP's analyse the behaviour of the values $d_{i,j} = (1 + AO_{i,j})^{-1}$. $AO_{i,j} = AO(Y_i(t_j), P_n(t_j))$, as defined in (11). Thus, if the set $\{d_{i_0,j} : j = 1, \ldots T\}$ contains mostly low values, the curve Y_{i_0} is a potential shift outlier; however, if it contains strong variations Y_{i_0} is a potential shape outlier.

To implement this idea, the authors of [4] measure the variations of the values in $\{d_{i_0,j} : j = 1, \ldots T\}$ by comparing the MFSPD with the arithmetic means of the $(1 + AO)$'s, and so taking advantage of the fact that the MFSPD is the inverse of the harmonic means of the $(1 + AO)$'s. While the basic idea is very attractive to us, this implementation seems a bit too unnecessarily involved. Therefore, we propose to employ a simpler, and more usual, measure of variation: we compare the arithmetical means of the AO's with a modification of the standard deviations of those quantities. We denote by CSP_a (respectively CSPD_d) to a CSP computed with the methodology proposed by the authors (resp. in this discussion).

The use of the standard deviation comes from it being a non robust dispersion measure, as the aim is to highlight variations in the values $AO_{i_0,j}$'s when i_0 is fixed. The modification introduced here consists in considering the positive part of the distances between those quantities and their median, instead of considering the standard deviations directly. This is due to the fact that we are only interested in the largest values of the standard deviations and this modification highlights those values in the CSP_d. Being a bit more precise, a CSP_d is constructed as follows

1. For every $i = 1, \ldots, n$ compute the values

$$\begin{aligned} AO_{i,j} &= AO(Y_i(t_j), P_n(t_j)), j = 1, \ldots, T, \\ \overline{AO}_i &= \frac{1}{T} \sum_j AO_{i,j} \\ s_i^2 &= \frac{1}{T} \sum_j AO_{i,j}^2 - \overline{AO}_i^2 \end{aligned}$$

2. Compute the median of the quantities $s_1, \ldots, s_n$, m, and the values

$$s_i^+ = \max(s_i - m, 0), i = 1, \ldots, n.$$

3. Plot the pairs $(\overline{AO}_i, s_i^+), i = 1, \ldots, n$.

This procedure has the additional advantage of not involving depth functions. While (as rightly pointed out in the paper) a low depth does not guarantee an outlier, a high value of $\overline{AO}_i$ is a strong hint of outlierness, because this quantity is the mean of the outlierness of the values of $Y_{i_0}(t_j)$ in the

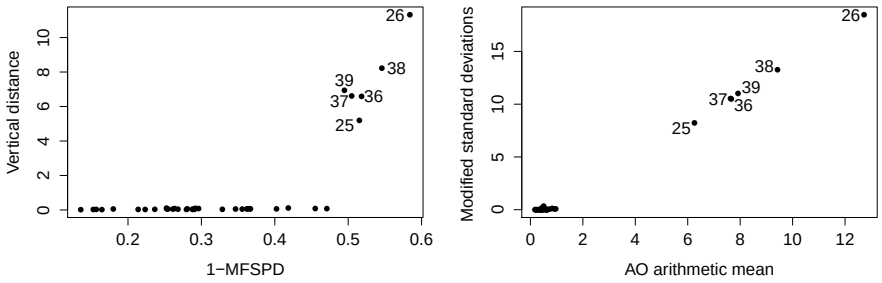

Fig. 4 CSP_a (left) and CSP_d (right) of the octane data

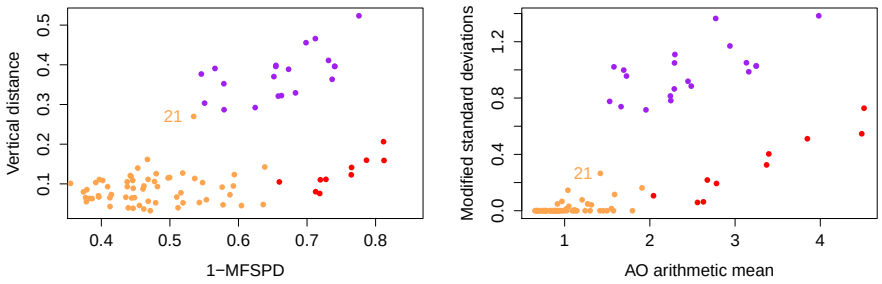

Fig. 5 CSP_a (left) and CSP_d (right) of the tablets data

sets $\{Y_i(t_j), i = 1, \ldots, n\}, j = 1, \ldots T$. Obviously the dimension of the curves affects the computation of the CSP_d in the same way as it does to the CSP_a.

The obtained results are encouraging as, in the examples considered, the CSP_d's give slightly better results than those produced by the CSP_a's. Next we present the result of applying the previous procedure to the octane and tablets datasets; there the curves are, respectively, one and three dimensional. Since the CSP_a of the octane data is not included in the paper, in Figure 4 we present the CSP_a (left) and the CSP_d (right) of this data set. Both two CSP's are similar, but the separation between the outlying points and the rest is proportionally greater in the CSP_d.

Concerning the tablets, the set the authors handle is composed by all the 90mg tablets plus a random sample taken from the 250mg tablets, thus, we have decided to randomly construct a similar set (which, of course, is going to be different to the one handled by the authors) and, consequently, in Figure 5 we show the CSP_a and the CSP_d of the data we have obtained. Again, although both CSP's are similar, the position of the point 21 in the 90mg tablets is quite different, being declared as outlier with the CSP_a but not with the CSP_d.

In conclusion, we can say that both methodologies produce similar graphs, with no too big differences between them; but the one based on the mean and standard deviation seems more natural to us and, in the analysed examples, produces slightly better results.

References

1. Cuesta-Albertos, J. A. and Nieto-Reyes, A., The random Tukey depth, Computational Statistics & Data Analysis, 52, 4979–4988 (2008)
2. Cuesta-Albertos, J. A. and Nieto-Reyes, A., The Tukey and the random Tukey depths characterize discrete distributions, Journal of Multivariate Analysis, 99, 2304–2311(2008)
3. Cuesta-Albertos, J. A. and Nieto-Reyes, A., Functional Classification and the Random Tukey Depth. Practical Issues, In Combining Soft Computing and Statistical Methods in Data Analysis, Advances in Intelligent and Soft Computing Series, Editors: Borgelt, Christian and Gonzlez-Rodrguez, Gil and Trutschnig, Wolfgang and Lubiano, Mara and Gil, Mara and Grzegorzewski, Przemyslaw and Hryniewicz, Olgierd, Springer Berlin / Heidelberg, 77, 123-130 (2010)
4. Hubert, M., Rousseeuw, P. and Segaert, P., Multivariate Functional Outlier Detection, Statistical Methods and Applications (2015)
5. Liu, X. and Zuo, Y., Computing projection depth and its associated estimators, Statistics and Computing, 24, 51–63 (2014)
6. Mosler, K. and Hoberg, R., Data analysis and classification with the zonoid depth, In Data depth: robust multivariate analysis, computational geometry and applications, DIMACS Ser. Discrete Math. Theoret. Comput. Sci., Amer. Math. Soc., 72, 49–59 (2006)
7. Rasmussen, Carl Edward and Williams, Christopher K. I., Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning), The MIT Press (2005)